Continuing Evolution of *Burkholderia mallei* Through Genome Reduction and Large-Scale Rearrangements

Liliana Losada*†,[1,2], Catherine M. Ronning†,[1], David DeShazer[3], Donald Woods[4], Natalie Fedorova[1], H. Stanley Kim[5], Svetlana A. Shabalina[6], Talima R. Pearson[7], Lauren Brinkac[1], Patrick Tan[8,9], Tannistha Nandi[8], Jonathan Crabtree[10], Jonathan Badger[11], Steve Beckstrom-Sternberg[7], Muhammad Saqib[12,13], Steven E. Schutzer[14], Paul Keim[7], and William C. Nierman[1,15]

[1]J. Craig Venter Institute, Rockville, Maryland

[2]Trinity University, Washington, DC

[3]U.S. Army Medical Research Institute of Infectious Diseases, Fort Detrick, Maryland

[4]Department of Microbiology and Infectious Diseases, University of Calgary, Calgary, Alberta, Canada

[5]Bioinformatics and Functional Genomics Laboratory, College of Medicine, Korea University, Seoul

[6]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland

[7]Center for Microbial Genetics and Genomics, Northern Arizona University

[8]Genome Institute of Singapore, Singapore, Singapore

[9]Duke-National University of Singapore Graduate Medical School, Durham, North Carolina

[10]Bioinformatics Resource Center, University of Maryland Baltimore County

[11]J. Craig Venter Institute, San Diego, California

[12]Veterinary Research Center, Barka, Sultanate of Oman

[13]University of Agriculture, Faisalabad, Pakistan

[14]Department of Medicine, University of Medicine and Dentistry—New Jersey Medical School

[15]Department of Biochemistry and Molecular Biology, The George Washington University School of Medicine

†These authors contributed equally to this work.

*Corresponding author: E-mail: llosada@jcvi.org.

Accepted: 19 January 2010 **Associate editor:** William Martin

Abstract

Burkholderia mallei (Bm), the causative agent of the predominately equine disease glanders, is a genetically uniform species that is very closely related to the much more diverse species *Burkholderia pseudomallei* (Bp), an opportunistic human pathogen and the primary cause of melioidosis. To gain insight into the relative lack of genetic diversity within Bm, we performed whole-genome comparative analysis of seven Bm strains and contrasted these with eight Bp strains. The Bm core genome (shared by all seven strains) is smaller in size than that of Bp, but the inverse is true for the variable gene sets that are distributed across strains. Interestingly, the biological roles of the Bm variable gene sets are much more homogeneous than those of Bp. The Bm variable genes are found mostly in contiguous regions flanked by insertion sequence (IS) elements, which appear to mediate excision and subsequent elimination of groups of genes that are under reduced selection in the mammalian host. The analysis suggests that the Bm genome continues to evolve through random IS-mediated recombination events, and differences in gene content may contribute to differences in virulence observed among Bm strains. The results are consistent with the view that Bm recently evolved from a single strain of Bp upon introduction into an animal host followed by expansion of IS elements, prophage elimination, and genome rearrangements and reduction mediated by homologous recombination across IS elements.

Key words: bacterial evolution, comparative genomics, genome erosion, bacterial virulence.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
22 JAN 2010	**N/A**	-

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Continuing evolution of Burkholderia mallei through genome reduction and large-scale rearrangements. J Genome Biol Evol 2010:103-116	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
Losada, L Ronning, CM DeShazer, D Woods, D Fedorova, N Kim, HS Shabalina, SA Pearson, TR Brinkac, L Tan, P Nandi, T Crabtree, J Badger, J Beckstrom-Sternberg, S Saqib, M Schutzer, SE Keim, P Nierman, WC	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
United States Army Medical Research Institute of Infectious Diseases, Fort Detrick, MD	**TR-09-131**
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT
Burkholderia mallei (Bm), the causative agent of the predominately equine disease glanders, is a genetically uniform species that is very closely related to the much more diverse species Burkholderia pseudomallei (Bp), an opportunistic human pathogen and the primary cause of melioidosis. To gain insight into the relative lack of genetic diversity within Bm, we performed whole-genome comparative analysis of seven Bm strains and contrasted these with eight Bp strains. The Bm core genome (shared by all seven strains) is smaller in size than that of Bp, but the inverse is true for the variable gene sets that are distributed across strains. Interestingly, the biological roles of the Bm variable gene sets are much more homogeneous than those of Bp. The Bm variable genes are found mostly in contiguous regions flanked by insertion sequence (IS) elements, which appear to mediate excision and subsequent elimination of groups of genes that are under reduced selection in the mammalian host. The analysis suggests that the Bm genome continues to evolve through random IS-mediated recombination events, and differences in gene content may contribute to differences in virulence observed among Bm strains. The results are consistent with the view that Bm recently evolved from a single strain of Bp upon introduction into an animal host followed by expansion of IS elements, prophage elimination, and genome rearrangements and reduction mediated by homologous recombination across IS elements.

15. SUBJECT TERMS
Burkholderia mallei, glanders, evolution, genome reduction, large-scale rearrangement, comparative genomics, genome erosion, bacterial virulence

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**SAR**	**15**	

Introduction

Burkholderia mallei (Bm) is a pathogen that is not found outside its mammalian host (Sanford 1995), yet its genome is highly similar to that of *Burkholderia pseudomallei* (Bp), a versatile, saprophytic pathogen endemic to the warm, wet soils of South East Asia and Northern Australia (Dance 1991). Bm causes glanders in equids, usually resulting in chronic infections but can cause fatal, acute infection in humans and other domesticated mammals. Its historical use as a biological weapon has led the Centers for Disease Control and prevention to classify Bm and Bp as category B select agents. Bp causes the human disease melioidosis and has been associated with disease in numerous hosts beyond mammals, including birds, reptiles, and even survival inside amoeba (Inglis et al. 2000).

It has been suggested that Bm evolved from a single strain of Bp, after an ancestral strain infected an animal host and then lost genes not required for survival in the host, ultimately becoming an obligate pathogen (Godoy et al. 2003; Nierman et al. 2004). This hypothesis is supported by the genomic similarity shared by two reference strains: both Bp K96243 and Bm ATCC23344 possess two circular chromosomes, nearly all Bm genes have orthologs in Bp, and Bp has roughly 1,200 additional genes. The versatility of Bp's host range and living environments is reflected in the species' genome. For example, there exist a wide array of genomic islands (GIs) variably represented across different Bp strains that give each strain different characteristics (Sim et al. 2008; Tuanyok et al. 2008; Tumapa et al. 2008). Moreover, it is proposed that these GIs were acquired via horizontal gene transfer from other soil saprophytes, consistent with a life in diverse environments outside of a host. Lastly, different GIs are present in strains isolated from different regions of the world (Sim et al. 2008; Tuanyok et al. 2008; Tumapa et al. 2008), demonstrating that the genomes are adapted to different environmental conditions. In contrast, the underlying mechanism for host and environmental restriction in Bm is not clearly understood.

These observations are similar to those in other bacterial genera where a "host-generalist" pathogen (in this case Bp) has undergone genome erosion (Ochman and Davalos 2006) that resulted in a "host-restricted" pathogen (Bm). Bm appears to be in an intermediate stage of erosion similar to *Shigella flexneri*, *Salmonella typhi*, *Francisella tularensis* (Ochman and Davalos 2006). Genome evolution in bacterial pathogens is a dynamic process that can occur over long periods of time but also during the span of short infections in a host (Oliver et al. 2000; Kraft et al. 2006). Under great selective pressures, such as survival in a host, unnecessary or deleterious genes could mutate rapidly or be lost entirely. Recombination across repeated sequences in a genome can lead to rapid gene mutation and loss. The genomes of Bp and Bm have very high contents of simple sequence repeats and IS elements that could have mediated recombination, resulting in the common gene disruptions, genomic inversions, translocations, duplications, and deletions observed in the reference Bm genome (Nierman et al. 2004). However, the extent of these gene losses and rearrangements across multiple Bm isolates has not been studied, and thus, it is unknown how common these events have been across the species.

We hypothesized that comparative genomic analysis of several Bm and Bp genomes would reveal a core set of genes essential for survival and virulence in a mammalian host, and elucidate genes involved in environmental survival. In addition, the analysis would also clarify the evolutionary process from a Bp ancestor to a modern Bm genome. Our results provide strong evidence for the evolution of Bm from a single ancestral Bp strain whose genome eroded through IS-mediated elimination of clusters of genes. The analysis suggests that the deleted genes were those that contributed to survival of Bp in the environment but were nonessential to the life of Bm as a mammalian pathogen. In addition, several clusters of genes were variably lost from different Bm strains, suggesting that the Bm genomes still contain genes that are under reduced selection in the equid host and might be unnecessary for survival in the host. Last, the results show that the Bm continues to undergo genomic erosion that can lead to reduced virulence.

Materials and Methods

Bacterial Strains Seven *B. mallei* strains and eight Bp strains were selected for sequencing and analysis based on geographic origin and virulence status (table 1).

Sequencing and Annotation The Bm type strain ATCC23344 was previously sequenced (Nierman et al. 2004). Three Bm strains (NCTC10229, NCTC10247, and SAVP1) were sequenced with full closure and manually annotated using approaches previously described (Nierman et al. 2004). The remaining three (2002721280, ATCC10399, and PRL-20) were sequenced to 8× Sanger sequence coverage by the whole-genome shotgun method (Fleischmann et al. 1995) without closure, assembled using Celera Assembler (Myers et al. 2000), and contigs oriented by alignment to the reference strain ATCC23344 using PROMER (Delcher et al. 2002). Open reading frames (ORFs) were predicted and annotated automatically using GLIMMER (Salzberg et al. 1998; Delcher et al. 1999). Pseudochromosomes were constructed from the ordered scaffolds, using manual examination where necessary. Bp strains 1106a, 1710b, and 668 were sequenced with full closure and manual annotation, whereas 1655, 406e, S13, and Pasteur 52237 were sequenced without closure and annotated automatically to 8× coverage. The Bp type strain K96243 was downloaded for analysis (Holden et al. 2004).

Table 1
Burkholderia mallei and *B. pseudomallei* Strains Used in This Study

	GenBank accession number	Virulent	Source	MLST	Size (bp)		Total genes	Variable genes (% of genome)[a]
					Chromosome I	Chromosome II		
B. mallei								
ATCC23344 Nierman et al. (2004)	NC_006348, NC_006349	Yes[b]	Burma 1944	40	3,510,148	2,325,379	5,229	1,773 (34%)
NCTC10229	NC_008836, NC_008835	Yes[b]	Hungary 1961	40	3,458,208	2,284,095	5,519	2,063 (37%)
NCTC10247	NC_009080, NC_009079	Attenuated[b]	Turkey 1960	100	3,495,687	2,352,693	5,869	2,413 (41%)
SAVP1	NC_008785, NC_008784	No	Schutzer et al. (2008)	40	3,497,479	1,734,922	5,200	1,744 (33%)
2002721280	NZ_AANX00000000[c]	No[b]	Pasteur Institute	40	—	—	5,300	2,239 (35%)
ATCC10399	NZ_AAHN00000000[c]	Yes[b]	China 1942	40	—	—	5,749	1,844 (40%)
PRL-20	NZ_AAZP00000000[c]	Yes	Pakistan 2005	40	—	—	5,469	2,013 (37%)
B. pseudomallei								
K96243 Holden et al. (2004)	NC_006350, NC_006351	Yes	Thailand 1996	10	4,074,542	3,173,005	6,324	688 (11%)
1106a	NC_009076, NC_009078	Yes	Thailand 1993	70	3,988,455	3,100,794	7,187	1551 (21%)
1710b	NC_007434, NC_007435	Yes	Thailand 1999	177	4,126,292	3,181,762	7,088	1452 (20%)
668	NC_009074, NC_009075	Yes	Australia 1995	129	3,912,947	3,127,456	7,232	1388 (19%)
1655	NZ_AAHR00000000[c]	Yes	Australia 2003	131	—	—	6,980	1344 (19%)
406e	NZ_AAMM00000000[c]	Yes	Thailand 1988	211	—	—	6,880	1244 (18%)
S13	NZ_AAHW00000000[c]	Yes	Singapore	51	—	—	7,217	1581 (22%)
Pasteur 52237	NZ_AAHV00000000[c]	Yes	Viet Nam	411	—	—	7,154	1518 (21%)

[a] Core genome is 3,456 genes for Bm and 5,636 for Bp.
[b] Virulence determined by Syrian hamster infection model. Three groups of female Syrian hamsters (five per group) were infected by the intraperitoneal route with a range of 10^1–10^3 cfu for each strain of *B. mallei* examined. Mortality was recorded daily for 14 days and on day 15, the surviving animals from each group were euthanized.
[c] WGS, whole-genome shotgun sequencing (unfinished).

Analysis of Functional Role Categories Proportions of genes in each functional role category were calculated for each strain and then averaged over all seven Bm strains, over four virulent Bm strains, or over three avirulent Bm strains. *T*-tests were performed on the square root transformed percentage data to determine the significance of the difference between core and variable genes.

Identification of Shared and Strain-Specific Genes Coding sequences (CDSs) from each strain were aligned against the whole-genome sequence of every other strain using the Program to Assemble Spliced Alignments (Haas et al. 2003). All CDSs that could not be aligned were thus assumed to be specific to that strain relative to the strain against which it was aligned.

Identification of Paralogs CD-Hit was used to identify paralogs with 90% amino acid sequence identity within each of the Bm genomes.

Pan-Genome Analysis The pan-genome analysis was carried out as described previously (Tettelin et al. 2005). Very briefly, after sequentially comparing the seven Bm strains and the eight Bp strains in all possible combinations, the size of the species core- and pan-genomes were extrapolated (for detailed statistical calculations, see Tettelin et al. 2005). The core genome analysis was also conducted using OrthoMCL with a Blast *e* value cutoff of 1×10^{-5} and an inflation parameter of 1.5. The OrthoMCL output was used to construct tables of shared orthologs and strain-specific genes.

Whole-Genome Alignments WebACT (Abbott et al. 2005) and the multigenome homology tools at the Pathema web site (http://pathema.jcvi.org) were used to generate alignment images with *e* value cutoff of 1×10^{-5}.

Construction of Species Tree First, orthologous proteins (60–80% identical over at least 90% of their length) from Bm ATCC23344, Bp K96243, *B. thailandensis* E264, and *B. cenocepacia* AU 1065 were identified by cluster analysis. From this set, all proteins annotated as "putative," "domain," "family," and "related," as well as all hypothetical and unknown proteins, were eliminated. The selected proteins from each of the four species were concatenated and searched individually against the complete protein sets of *B. ambifaria* MC40-6, *B. cepacia* AMMD, *B. multivorans* ATCC17616, *B. phymatum* STM815, *B. phytofirmans* PsJN, *B. vietnamiensis* G4, *B. xenovorans* LB400, and *Pseudomonas aeruginosa* PA7 using BlastP to identify orthologs from

these species. The final set, which consisted of 56 proteins from each of the 12 species that were 60–80% identical over at least 95% of their length, were aligned using Muscle (Edgar 2004) then concatenated (supplementary table 1, Supplementary Material online). Bootstrapped maximum likelihood trees were calculated from the concatenated protein set using the PHYLIP package applying the JTT substitution model with a gamma distribution ($\alpha = 0.5$) of rates over four categories of variable sites, and a consensus tree was produced from the bootstrap replicates. Bootstrapped maximum parsimony and Neighbor-Joining trees were also created by PHYLIP, using the default parameters for those methods.

Identification of Orthologous Genes and Evolutionary Comparisons (dN/dS Analysis) Orthologous gene pairs were compiled from eight Bm strains by identifying symmetrical best hits between proteins from the reference strain ATCC23344 and the other seven Bm genomes using BlastP (http://www.ncbi.nlm.nih.gov/BLAST/) with a cutoff of 1×10^{-10}. Nucleotide sequence alignments were produced for orthologous pairs of ATCC23344 and each other Bm strain using MUSCLE and OWEN (Ogurtsov et al. 2002; Edgar 2004). Alignments of CDSs were guided by their corresponding amino acid sequence alignments (Kondrashov and Shabalina 2002). In cases where greater than 30% of the gaps or annotated regions of putative orthologs did not align or where pairs of sequences aligned perfectly (100% similarity), the sequence pairs were removed from further analysis. dN and dS values were calculated by Nei–Gojobori method (Nei and Gojobori 1986; Yang 1997). Overall, 1,018 and 219 detailed alignments were generated from the original 4,197 core and 996 variable Bm genes, respectively, and dN/dS ratios were estimated. Differences between rates of synonymous (dS) and nonsynonymous (dN) substitutions in the variable and core coding regions were analyzed with the Wilcoxon rank sum test.

Results

Genome Features Bm was reported to have evolved from a single strain of Bp that became highly adapted to its mammalian host (Godoy et al. 2003). In order to determine whether Bm was the result of genome reduction and clarify the mechanism of the proposed host adaptation, six Bm strains and seven Bp strain were sequenced and used in whole-genome comparative analyses. Each of the strains sequenced was selected based on their geographical or clinical isolation (table 1). Among the Bm strains, two were avirulent in a Syrian hamster model (SAVP1 and 2002721280) and one had reduced virulence (NCTC10247). The genome sizes of the seven sequenced Bm strains were similar, averaging 5.7 Mb (table 1). However, chromosome II of strain SAVP1 was significantly smaller than the other fully sequenced strains. The eight sequenced Bp strains averaged 7.2 Mb, approximately 1.5 Mb larger than that of Bm, and the corresponding chromosomes of the four fully sequenced and closed strains were relatively similar in size.

The genomic diversity among seven housekeeping genes in Bm and Bp strains was studied using multilocus sequence type (MLST) analysis (Maiden et al. 1998). Despite the differences in geographical distribution or virulence, all but one of the Bm strains belonged to the same MLST (http://bpseudomallei.mlst.net/; table 1), suggesting a lack of genetic diversity. The two identified MLST groups differed only in one nucleotide within the *gltB* locus, further demonstrating a highly similar genetic landscape. These results were consistent with Chantratita et al. (2006) who found that 21 isolates of Bm belonged to only one MLST type. In contrast, each of the eight Bp strains belonged to a different MLST group (table 1), and none of the Bp MLST groups matched the Bm MLST groups. Based on MLST relatedness, K96243 is the closest sequenced Bp relative, although there exist several Bp isolates with closer MLST profiles whose genome sequence is not known (Godoy et al. 2003). Combined, the genome properties and MLST data provide evidence for the clonal evolution of Bm from a single Bp ancestor.

Bm Lost Large Clusters of Bp Genes Associated with Environmental Survival To better understand the genome reduction among Bm strains, we performed reciprocal comparisons of all CDSs of each strain of one species with the genome sequence of each strain of the other species. The results showed that, as expected, many genes were Bp-specific relative to Bm (ranging from 1,122 to 1,488), whereas only very few (0–8) Bm-specific genes exist (data not shown). All the Bm-specific genes were either hypothetical proteins or phage integrases, presumably relics from a Bp ancestor. Interestingly, roughly 40% of the Bp-specific genes were clustered in the Bp genome and mapped to the GIs identified previously (Holden et al. 2004; Tuanyok et al. 2008; Tumapa et al. 2008; fig. 1). Furthermore, none of the GIs from the sequenced *B. pseudomallei* genomes are represented in any of the Bm genomes (data not shown). Almost all the remaining 60% of Bp-specific genes also clustered in the genome (fig. 1) and, in some cases, were deletions surrounding the GIs, similar to the observation made in a wide panel of Bp isolates (Sim et al. 2008). The loss of these GIs could explain why Bm is not found in the environment because many of the GIs lost in Bm have functions associated survival and competition in the soil environment (Holden et al. 2004; Tuanyok et al. 2008; Tumapa et al. 2008). For instance, at least four of the GIs lost encode for multidrug resistance pumps. In addition, several of the Bp GI encode for secondary metabolite clusters that could act as antibacteriacidals or antifungals (Duerkop et al. 2009), and thus allow Bp to compete in the soil, whereas Bm would be at a disadvantage.

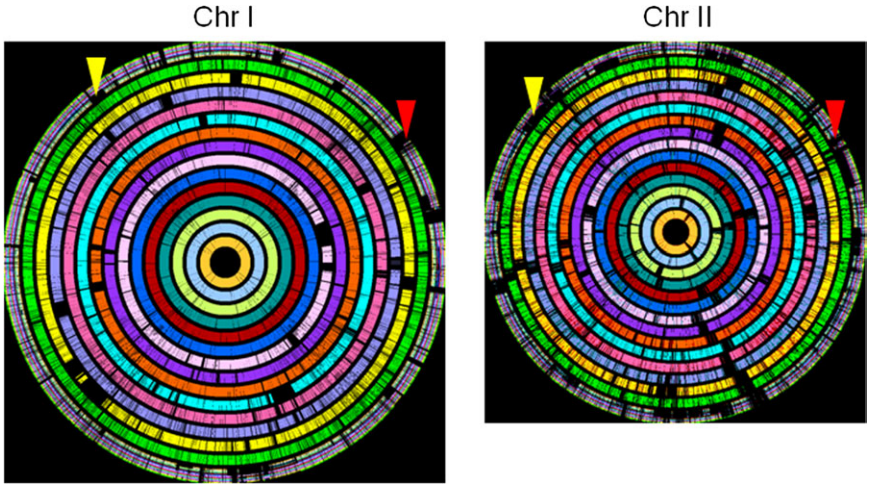

Fig. 1.—Multigenome alignment of eight Bp and seven Bm strains. Each circle represents a genome as presented in Materials and Methods. All genomes are aligned with Bp K96234 genome as a reference, which appears as the outermost multicolored circle. The Bp genomes are the eight outermost circles, and Bm genomes are internal. Areas in each color represent homologies between the subject genome and the reference. Areas in black in the reference chromosome (outermost circle) are regions present in K96243 but absent in query genome. Areas in black in each of the concentric circles are regions present in the query genome but absent from K96243. Representative Bp GIs are shown with red arrows. Representative clusters of Bp-specific genes absent from all Bm genomes (black on the K96243 ring) are highlighted with a yellow arrow.

Since the GI are proposed to be the source of environmental variability in Bp (Holden et al. 2004; Tuanyok et al. 2008; Tumapa et al. 2008), and given that these are absent from the Bm genome, we hypothesized that the entire genetic complement, or pan-genome (Tettelin et al. 2005), of Bm would be significantly reduced compared with Bp. Pan-genome analysis confirmed that the Bm strains were remarkably homogeneous in their gene content. The number of new genes dropped off precipitously and essentially leveled off after inclusion of only five genomes, indicating that sequencing additional Bm strains would not reveal a significant number of novel genes (fig. 2*A*). In other words, essentially all Bm genes will be identified after only 4 or 5 additional genomes are sequenced. In contrast, the number of new genes leveled off much more gradually in Bp (fig. 2*B*), suggesting that 25–50 new genes will be revealed with each newly sequenced strain.

Bm Has a Distinct Variable and Core Genome The Bm core genome is defined as the set of genes that is common to all strains, whereas the strain-specific variable gene sets contain genes that are absent from at least one of the other Bm genomes. The Bm core genome consisted of 3,456 genes, whereas the pan-genome (the core gene set plus variable genes) contained about 2,300 more (roughly 5,700 genes; table 1). SAVP1 and ATCC23344 had the fewest number of variable genes (1,773 and 1,774, respectively), whereas NCTC10247 had the most (2,413). Many of the core genes had duplicates and paralogs that were considered part of the variable gene set. The total number of duplicates or paralogs in each strain ranged from 240 to 253, most of which were annotated as IS elements. Consistent with the hypothesis that Bm evolved from a single strain of Bp, these Bm variable genes all had orthologs in Bp K96243, suggesting that the mammalian host environment offered no opportunity for new gene influx into the Bm pan-genome. These results suggest that the Bm pan-genome is closed and that the organism has entered an evolutionary bottleneck in the host.

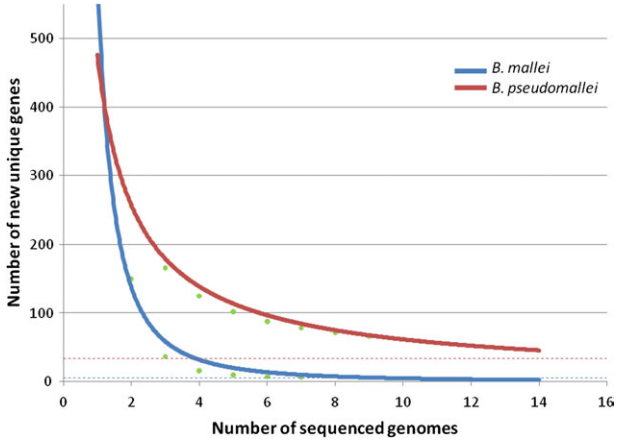

Fig. 2.—Pan-genome analysis of seven Bm and eight Bp strains. The CDSs in all Bm genomes (blue line) and Bp genomes (red line) were compared, and the number of new genes was plotted against the number of genomes used. The blue dashed line represents the extrapolated number of Bm strain-specific genes. The red dashed line represents the extrapolated minimum number of new genes discovered with each Bp genome.

Table 2
Percentages of Total Variable Genes within Each Functional Role Category

	Burkholderia pseudomallei		*B. mallei*	
Role category	**Mean (%)**	**Standard deviation (%)**	**Mean (%)**	**Standard deviation (%)**
Amino acid biosynthesis	1.49	2.37	2.16	0.89
Biosynthesis of cofactors, prosthetic groups, and carriers	0.83	1.55	1.02	0.16
Cell envelope	6.80	4.78	11.46	2.44
Cellular processes	6.12	3.44	12.42	3.16
Central intermediary metabolism	2.10	3.12	2.67	0.35
DNA metabolism	24.51	14.39	0.90	0.21
Energy metabolism	3.57	4.83	14.42	1.86
Fatty acid and phospholipid metabolism	0.66	1.33	3.89	0.78
Mobile and extrachromosomal element functions	29.34	14.70	0.82	0.41
Protein fate	3.56	4.18	6.56	1.26
Protein synthesis	0.76	2.14	1.09	0.41
Purines, pyrimidines, nucleosides, and nucleotides	0.00	0.00	1.85	0.78
Regulatory functions	8.84	5.75	16.78	1.23
Signal transduction	0.00	0.00	5.37	3.50
Transcription	2.36	2.39	0.75	0.29
Transport and binding proteins	7.33	6.52	17.85	1.83
Viral functions	1.73	3.97	0.00	0.00

NOTE.—Mean, standard deviation, and range are given for eight Bp strains and seven Bm strains. Hypothetical and unknown proteins and proteins of unknown function have been excluded.

The Bp core- and pan-genomes (ca. 5,300 and 7,500 genes, respectively; table 1) were larger than those of Bm. The variable genome of Bp ranged from 454 to 837, genes many of which were encoded within GIs. Interestingly, the variable genome in Bm encompassed a larger portion of the genome (33–41%) than in Bp (20.6%), suggesting that even with a relatively narrow genetic base, the genome of Bm is continuing to change, albeit without actual gain of genes to the pan-genome.

It is possible that the large Bm variable genome is an artifact of in vitro culture deletions that led to a loss of virulence (Schutzer et al. 2008). In vitro culture would remove the selective pressure on genes essential for survival in the mammalian host, leading to the loss of some of these genes. To address this possibility, the analysis was repeated after removing the two avirulent strains (SAVP1 and 2002721280). The size of the variable genome decreased by 610 genes, and accordingly, the core genome increased by 610 genes because those genes were shared among the remaining five strains. Interestingly, none of the 610 genes were lost from both avirulent strains showing that there exist at least two independent traits that are essential for virulence in a mammalian host (see below).

Analysis of functional role categories of variable genes among strains of Bm and of Bp revealed significant differences between the two variable genomes (table 2) that were consistent with each species life style. Much of the Bp variable genome was associated with phage elements or complete prophage (Ronning CM, Nierman WC, Ulrich RL, DeShazer D, in preparation) and had predominate gene functions of mobile and extrachromosomal elements (29.3%) and DNA metabolism (24.5%; table 2). These genes were probably acquired through lateral gene transfer in the soil environment. In contrast, the predominant roles in the Bm variable genes are cell envelope, cellular processes, energy metabolism, regulatory functions, and transport and binding (table 2). These functions are probably essential for survival and competition in the environment but are under lower selection in the host (Casadevall 2008).

Bm Variable Genes Exist in Multigene Contiguous Clusters Flanked by IS Elements For all Bm strains, the vast majority of the genes that were present in a particular strain but absent from one or more of the other strains tended to occur in contiguous clusters within that strain, with the total number of these variable gene clusters ranging from 9 to 18 for each strain (table 3). The presence or absence of these variable regions appeared to be the primary difference between Bm strains. In all strains, there were more variable gene clusters on chromosome II than chromosome I, even though chromosome II is smaller. The variable clusters among the seven Bm strains were classified into 24 groups based on sequence homology (table 3). The number of strains from which each cluster was absent ranged from 1 (clusters A, D, F, G, I, J, L, M, N, P, Q, and R) to 5 (cluster X). The variable regions varied greatly in size, from ~3.4 kb (cluster N) to ~269 kb (cluster Q).

Most of these clusters were flanked by transposases associated with IS elements, usually of the same type; however, a few were bounded by a transposase on one end only (table 3). Interestingly, some of these variable regions appeared contiguously in some genomes, for example,

Table 3
Variable Gene clusters in Bm

	5′ end	3′ end	Size (bp)	Boundary (5′/3′)	ATCC 23344	SAV P1	102 99	102 47	103 99	2002721 280	PRL-20	Number of putative virulence genes[a]	NRPS/PKS/ Multidrug efflux pump[b]
A	600,776	612,728	11,953	IS407A	X		X	X	X	X	X	1	
B	1,000,692	1,080,040	79,349	IS407A	X				X	X	X	11	RND
C	1,269,317	1,277,504	8,188	IS407A	X	X	X	X	X	X		4	
D	2,053,557	2,070,428	16,872	IS407A	X	X	X	X	X		X	5	
E	2,335,045	2,354,063	19,019	IS407A	X	X		X	X	X	X	2	PKS
F	2,527,011	2,629,142	102,132	ISBm2/IS407A	X	X	X	X	X		X	20	
G	3,320,410	3,346,619	26,210	ISBm2	X	X	X	X	X		X	6	
H	**104,657**	**170,441**	**65,785**	**IS407A**	**X**	**X**		**X**	**X**	**X**		**13**	**RND**
I	**173,242**	**319,417**	**146,176**	**ISBm2**	**X**	**X**	**X**	**X**	**X**		**X**	**32**	
J	**409,775**	**432,884**	**23,110**	**IS407A**	**X**		**X**	**X**	**X**	**X**	**X**	**5**	
K	**567,683**	**655,441**	**87,759**	**IS407A**	**X**		**X**	**X**	**X**	**X**		**15**	
L	**658,191**	**733,816**	**75,626**	**IS407A**	**X**		**X**	**X**	**X**	**X**	**X**	**9**	**RND**
M	**839,856**	**869,581**	**29,726**	**IS407A**	**X**		**X**	**X**	**X**	**X**	**X**	**6**	
N	**895,207**	**898,647**	**3,441**	**ISBm2**	**X**		**X**	**X**	**X**	**X**	**X**	**1**	
O	**1,015,758**	**1,061,756**	**45,999**	**IS407A**	**X**		**X**	**X**	**X**	**X**		**6**	**RND, PKS**
P	**1,176,775**	**1,225,744**	**48,970**	**ISBm1/IS407A**	**X**		**X**	**X**	**X**	**X**	**X**	**13**	**NRPS**
Q	**1,518,817**	**1,790,695**	**271,879**	**IS407A**	**X**		**X**	**X**	**X**	**X**	**X**	**64**	**NRPS, PKS**
R	**2,158,811**	**2,265,535**	**106,725**	**ISBm2**	**X**	**X**	**X**	**X**	**X**		**X**	**28**	**PKS**
S	1,136,910	1,145,707	8,798	None/IS407A		X	X	X		X	X	2	
T	**783,963**	**817,798**	**33,836**	**IS407A/ transposase**		**X**	**X**	**X**	**X**	**X**		**18**	
U	**2,650,189**	**2,695,429**	**45,241**	**IS407A**		**X**	**X**	**X**		**X**		**0**	
V	947,304	951,928	4,625	IS407A/none			X	X		X		1	
W	1,809,469	1,823,849	14,381	A, transposase OrfB/IS407A				X	X	X		0	
X	**1,237,829**	**1,245,922**	**8,094**	**IS407A/ISBma2**		**X**			**X**		**X**	**1**	**RND**

NOTE.—Each variable cluster was assigned a letter. Genomic locations for clusters A–R are from ATCC23344, where the bold font represents those located on chromosome II. Genomic locations for clusters S–W are from NCTC10247 (bold, chromosome II), and cluster X from NCTC10399 chromosome II. An X under each strain signifies that the cluster is presented in that genome.

[a] Virulence genes were determined by using MVirDB as described in Materials and Methods.

[b] NRPS, nonribosomal peptide synthase; PKS, polyketide synthase; RND, resistance nodulation-division like pump.

clusters C and T in SAVP1 and NCTC10229. By searching the sequence flanking the variable gene clusters against the strains from which the cluster is absent, putative excision points were mapped back to most of the strains (supplementary table 2, Supplementary Material online), which were invariably marked by transposases. In cases where the cluster was lost from several genomes, the excision point was the same in each genome. The results suggest that the variable gene clusters were present in the Bp ancestor and were differentially lost through IS element-mediated excision in different Bm strains.

Interestingly, several of the Bm variable genes had functions associated with survival and competition in a soil environment such as synthesis of secondary metabolites and drug resistance mechanisms. In total, 5/24 of the variable regions contain genes involved in nonribosomal peptide or polyketide synthesis (table 3). In addition, several metal ion resistance genes and stress-related proteins also belong to the variable gene set (data not shown). Lastly, a different set of five variable regions encode multidrug efflux pumps (table 3). Interestingly, genomes of NCTC10399 and SAVP1 encoded a 50-kb region containing a multidrug efflux pump that we had previously proposed as the source for aminoglycoside resistance (Nierman et al. 2004). Both of these genomes contained the same arrangement at the *amrAB-ompR* locus as Bp (data not shown; Moore et al. 1999) but contain a 6-bp deletion within the coding region of *amrB* that resulted in a two amino acid deletion in a highly conserved transmembrane motif (Putman et al. 2000) toward the C terminus of the protein. Both NCTC10247 and NCTC10299 contained a homolog of *amrA*, but the AmrB protein was truncated at amino acid 244 potentially resulting in sensitivity to aminoglycosides and macrolides. None of the remaining Bm genomes encoded for this region. The finding that this cluster is present in some Bm strains could help explain previous studies where a few of the Bm strains were resistant to aminoglycosides (Thibault et al. 2004). A recent study found several

aminoglycoside-sensitive clinical Bp isolates, some of which had also lost the entire *amrAB-ompR* locus, whereas others used an entirely different and unknown mechanism to repress expression of the operon (Trunck et al. 2009), suggesting that this locus is not necessary for survival in the host.

The Bm Genome Has Undergone a Dramatic Expansion of IS Elements That Mediated Extensive Intrachromosome Rearrangements within the Bm Strains

Whole-chromosome Rearrangements

In addition to the IS elements flanking the variable gene clusters, each Bm strain had a considerable repertoire of IS elements (ranging 166–218, supplementary fig. 1, Supplementary Material online). In particular, IS elements of the type IS407A had undergone a significant expansion in all the sequenced Bm strains, accounting for 76% of all IS elements (supplementary fig. 1, Supplementary Material online). Interestingly, most of the IS407A elements in Bm did not have the flanking 4-bp repeat that result from a transposon insertion (88% in chromosome I: 88% and 77% in chromosome II; DeShazer et al. 2001), suggesting that these elements had been subject to homologous recombination. Bias in base composition among the existing 4-bp repeats suggested that the initial transposon insertions within the chromosome were nonrandom, but rearrangement since then was random (fig. 3*A*). Whole-genome alignments demonstrated that Bm chromosomes were dramatically and extensively rearranged by recombination across IS407A elements (fig. 3*B*). Among the Bm strains, none of the IS407A rearrangements occurred between chromosome I and chromosome II. Intriguingly, Bp contained an average of seven IS elements per genome (supplementary fig. 1, Supplementary Material online), but these have not catalyzed such genome-wide rearrangements (fig. 3*C*). Thus, it is unclear whether there exist environmental selective pressures that maintain Bp's genomic arrangement, as in *Salmonella typhimurium* (Kothapalli et al. 2005) or whether rearrangements occur in Bm due to its high IS element content.

rrn Rearrangements in Bm

Chromosome I. It has been reported that many host-specific pathogens, so called specialists, have undergone considerable ribosomal RNA operon rearrangements when compared with their generalist relatives (Liu and Sanderson 1998). We investigated whether any of the large-scale rearrangements observed in Bm also affected the position and organization of *rrn* operons when compared with Bp. All the finished Bp strains (K96243, 1106a, 1710b, and 668) shared the same number, distribution, and organization of the *rrn* operons. There are three complete operons in chromosome I and one on chromosome II that all share the same order: *rrs*(16S)—*tRNA-Ile*—*tRNA-Ala*—*rrl*(23S)—*rrf*(5S). In contrast, in each of the four completely sequenced Bm strains, there were different numbers and distributions of the *rrn* loci (fig. 4*A*). Each chromosome I had at least one complete *rrn* locus with the same order as described above. However, each Bm strain had lost an entire operon from chromosome I, and the third locus was interrupted at the 1,427 bp of the 23S locus. SAVP1 had an additional remaining degenerate *rrn* locus on chromosome I that had an *IS*407A element interrupting the 23S locus at position 284. This IS element had the 4-bp repeat associated with an insertion event (DeShazer et al. 2001). Interestingly, the 5S locus was lost at all the degenerate locations. These results suggest that at least two 23S loci in Bm are susceptible to mutations via insertion of IS elements or phages that drive the loss of the 5S gene as well.

In addition to the loss of the 3′ sequence at two loci, each of the Bm strains displayed a different organization of *rrn* operons on chromosome I (fig. 4*A*). Despite a considerable degree of rearrangement, the orientation of the *rrn* operons was always in the direction of replication, consistent with observations in other species (Liu and Sanderson 1998; Shu et al. 2000). However, rearrangements in other species, like *Salmonella* and *Shigella*, almost always resulted in *rrn* operons that are equidistant from the origin of replication (Kothapalli et al. 2005). Compared with Bp, only ATCC23344 had an *rrn* operon at the same distance (0.2 Mb) from the origin of replication. Interestingly, NCTC10247 (reduced virulence strain) had a drastic rearrangement that left the *rrn* locus 1.1 Mb away from the origin of replication and also resulted in a chromosome with differently sized replichores (fig. 4*B*). It has been proposed that *rrn* loci must be close to the *ori* for adequate expression of ribosomal components necessary during cellular division (Schmid and Roth 1987; Kothapalli et al. 2005). In addition, *Escherichia coli* strains with differently sized replichores are at a growth disadvantage (Lesterlin et al. 2008). Thus, it is possible that the attenuation in virulence in NCTC10247 can be explained by these genomic constraints. The growth rate of NCTC10247 in rich media over a 24-h span was only slightly slower than NCTC10229 (average t_d = 191 min and 193 min, respectively). However, during early exponential growth, the doubling time (t_d) of NCTC10247 was 104 min compared with only 79 min for NCTC10229. In an animal host, this difference in growth could be sufficient to explain the attenuation.

Chromosome II. Neither of the *rrn* loci on Bm chromosome II have been subject to degeneration. When compared with Bp chromosome II, the additional *rrn* locus could be the result of intrachromosomal duplication of the existing locus or due to an exchange between the two chromosomes. Whole-genome alignments with Bp revealed that this locus was part of a 46-kb interchromosomal exchange between chromosome I and chromosome II flanked by IS407A

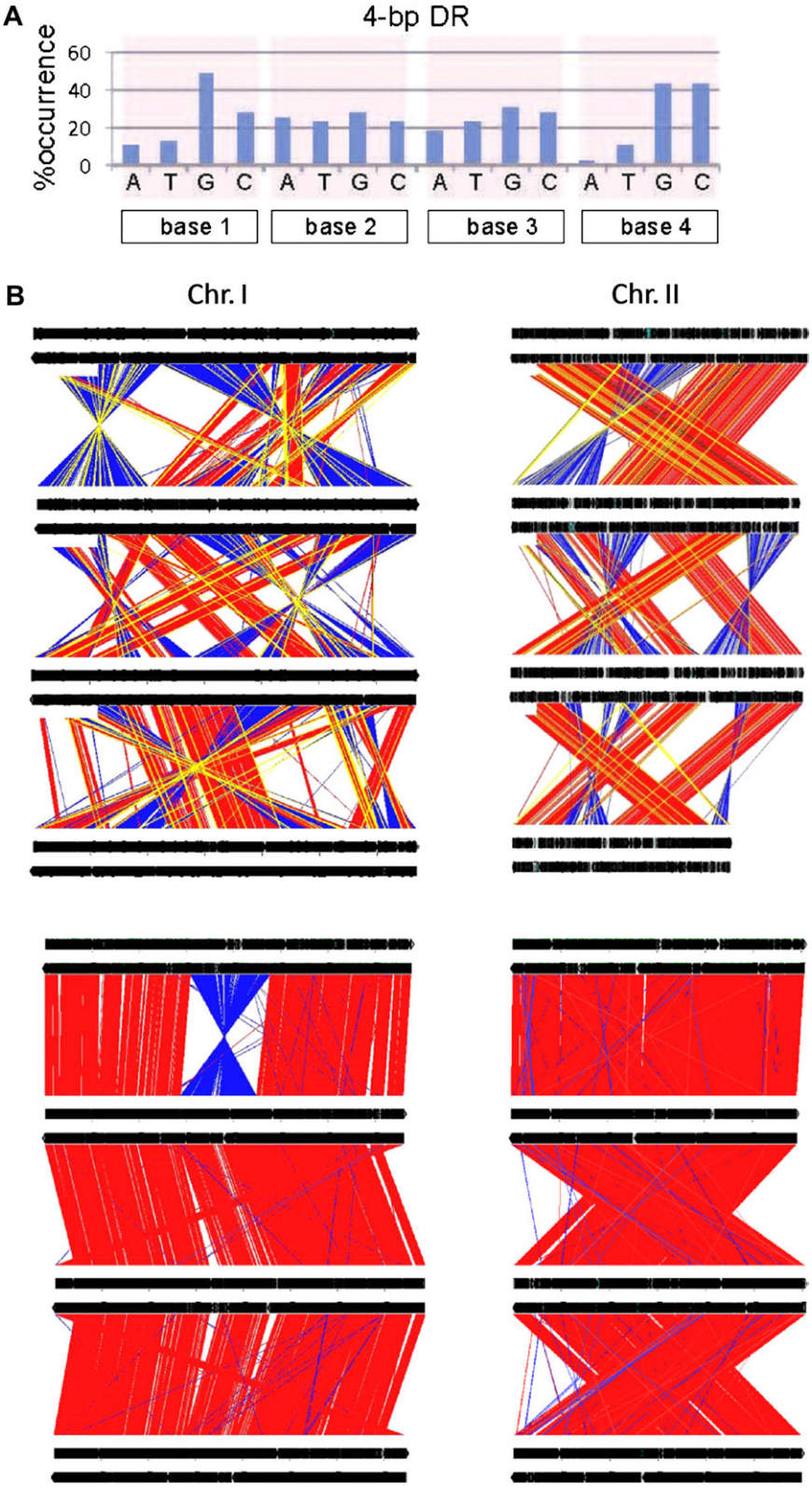

Fig. 3.—IS407A rearrangement of whole genomes. (A) Relative occurrence of the nucleotides in the 4-bp direct repeat of IS407A element insertion is shown as bar graphs for each position in the box below. (B) Four fully sequenced Bm genomes were aligned using WebACT. Red lines denote homology between chromosomes organized in the same orientation. Blue lines show homology but inverse orientation in each chromosome. Yellow lines show the presence of IS407A elements. Regions with no homology are shown by the absence of red or blue lines. (C) Four fully sequence Bp genomes were aligned, as described for Bm.

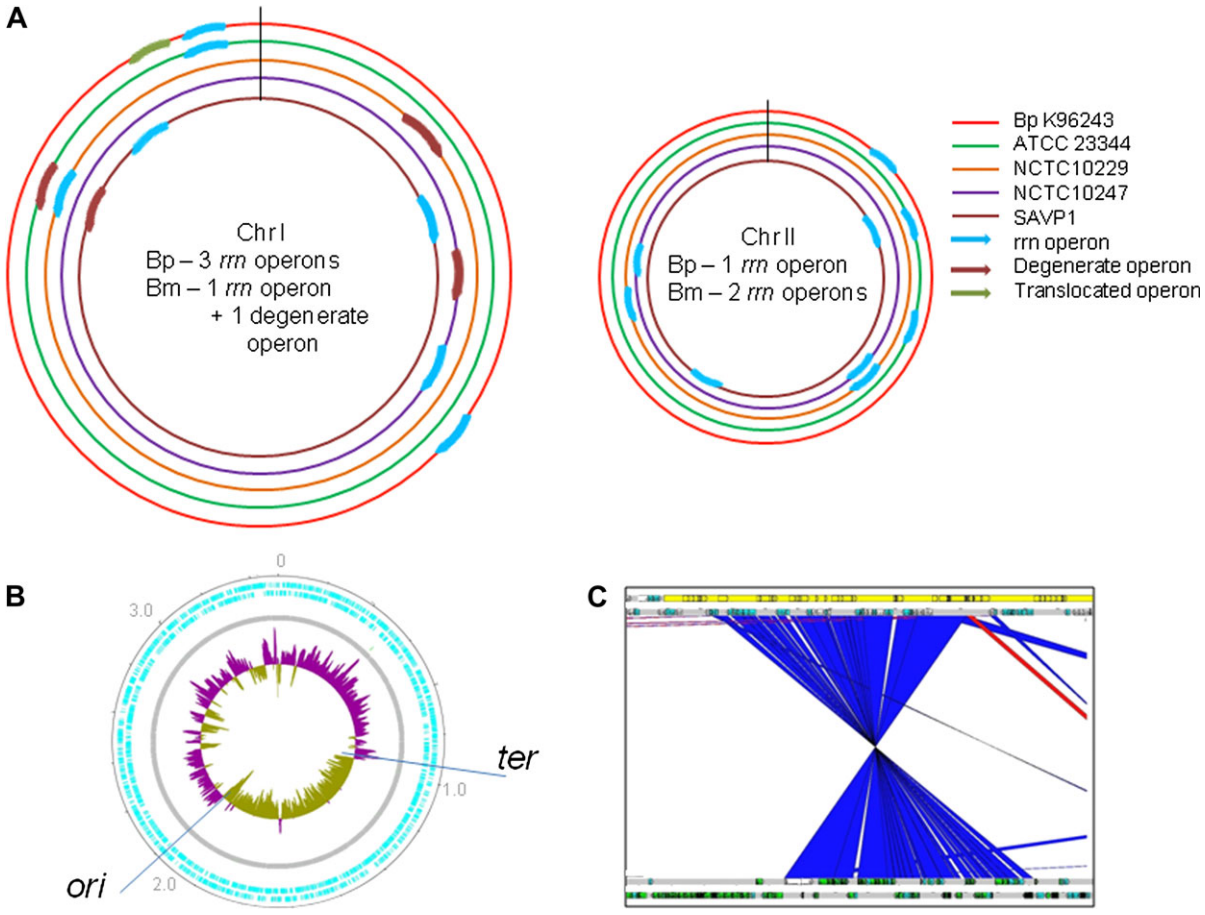

Fig. 4.—IS407A mediated rearrangements of *rrn* and replichores among Bm strains. (A) *rrn* rearrangements due to IS407A recombinations. The outermost ring corresponds to Bp K96243 but is a representative of all Bp genomes. Green, ATCC23344; orange, NCTC10229; purple, NCTC10247; brown, SAVP1. The brown *rrn* cluster represents the locus rearranged into chromosome II in Bm. Red bars represent degenerate *rrn* loci. (B) guanine/cytosine-skew representation of the NCTC10247 genome generated in DNAplotter (Carver et al. 2009). Green represents a negative guanine/cytosine-skew suggesting ORF are oriented in the negative strand and purple represents a positive guanine/cytosine-skew suggesting ORF oriented in the positive strand. The origin of replication for NCTC10247 chromosome I is predicted at around 2.3 Mb and the termination around 1.0 Mb. (C) Alignment of chromosome II of ATCC23344 with chromosome I of BpK96243 as Bp representative. Regions of homology are represented by blue color. For the sake of clarity, only the genomic regions of interest are depicted.

elements (fig. 4C). This exchange occurred after the divergence between Bp and Bm because all the Bp strains carried this cluster on chromosome I and all Bm stains on chromosome II. As in chromosome I, the organization of the *rrn* loci on Chromosome II was not conserved, and each Bm strain had a different organization around the chromosome. However, all loci are oriented in the same direction of transcription, with none being as close to the origin of replication as the *rrn* locus in the Bp chromosome II.

fliP IS407A Element Insertions

Bm is nonmotile, and thus, it was originally surprising to find that flagellar biosynthesis genes were present in the ATCC23344 genome with only one obvious mutation: an IS407A insertion into *fliP* (Nierman et al. 2004). Comparative analysis of the seven Bm strains showed that all the flagellar genes are present in all strains, but each one has an IS407A element at the same location in *fliP*, 124 bp from the start position. In all of these genomes, the N-terminal disruption of *fliP* also resulted in a 4-bp GACG complementary direct repeat that suggests the IS element was initially introduced via a transposition event. None of the Bp strains have a similar *fliP* mutation. These results suggest that functional flagella are necessary for environmental survival or generalist behavior but not for survival or virulence in the narrow Bm host range. Furthermore, the retention of all other flagellar genes in Bm suggests that those might be used as an alternate secretion apparatus similarly to *Buchnera* spp. (Toft and Fares 2008).

Interestingly, three different types of alleles were identified (fig. 5) among the seven strains. In NCTC10247, NCTC10229, and 2000721280, only the IS407A element

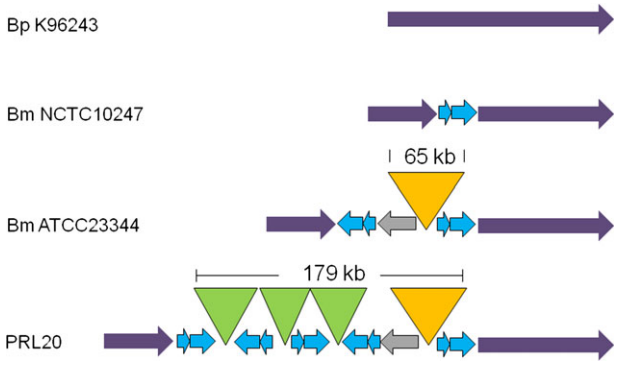

Fig. 5.—Genomic organization of the *fliP* locus in Bp and Bm. The wild-type *fliP* locus is present in all Bp. The *fliP* CDS is represented by dark purple rectangles. The NCTC10247 allele is interrupted by an IS407A (aquamarine) element. In ATCC23344, an ISBma1 (gray) is located upstream of the IS407A element and an additional 65 kb was inserted at this location. PRL20 had additional IS407A mediated insertions into *fliP*. Figures are not to scale, and IS407A elements in PRL-20 were made smaller.

disrupts the gene. In ATCC23344, SAVP1, and NCTC10399, an additional 65-kb region was located adjacent to the IS407A element. This region was flanked by phage-associated proteins on the end closest to *fliP*-C terminus and by an ISBma1 element closest of *fliP*-N terminus. In NCTC10247 and NCTC10299, this 65-kb genomic region is located elsewhere on chromosome I, flanked by ISBma and IS407A, suggesting that the arrangement in the latter group is due to a recombination event across the two IS407A elements, perhaps aided by the ISBma1 transposase. A third allele present in PRL-20 shares all the loci present in the ATCC23344 allele, but has at least three additional IS-element flanked insertions, that resulted in 179-kb insertion at *fliP*. None of the other IS407A elements within this region contain the 4-bp complementary direct repeat, further signifying that these insertions were not a result of transposition but instead due to intrachromosomal recombination mediated by IS407A. These observations suggest that genes in the Bm genome under no selection in the equid host will acquire IS insertions, and conversely, those genes without IS may be experiencing selection in the host.

Loss of Virulence Is Explained by IS-Mediated Loss of Essential Gene Clusters We wished to determine if any of the variable clusters contained virulence genes, particularly those absent from the two avirulent and one attenuated strain. Putative virulence genes were identified by blasting against the MvirDB database (Zhou et al. 2007; supplementary table 3, Supplementary Material online). Several of the clusters contained putative virulence genes, five of which (groups D, F, G, I, and R) were absent only from the avirulent strain 2002721280 and five (groups J, L, M, P, and Q) were absent solely from avirulent strain SAVP1. It has recently been reported that SAVP1 lacked the entire animal type III secretion system (TTSS) gene complex that was essential for virulence (Nierman et al. 2004; Ulrich and DeShazer 2004; Schutzer et al. 2008). The TTSS was encoded in the variable gene cluster P (table 3) that was lost through IS-mediated deletion in SAVP1 but was present in all other Bm stains. Because of its obvious virulence deficiency, no further analysis was done on this strain.

Analysis of the other avirulent strain did not immediately result in an obvious virulence defect. However, clusters D and F were lost through IS407A recombination. These clusters contain amino acid synthesis and transporters that probably resulted in a strain auxotrophic for lysine and ornithine and at least partially deficient in its capacity to uptake several amino acids (glutamate, aspartate, leucine, valine, and isolecuine). Indeed, 200272180 did not grow on minimal media (data not shown). Thus, it is likely that these deficiencies are sufficient to explain the lack of virulence observed in 2002721280, as was demonstrated for a branched-chain amino acid auxotroph of Bp (Atkins et al. 2002). Alternatively, the presence of large numbers of regulatory genes within the variable gene clusters lost from 2002171280 may, together with the identified virulence genes present within the clusters, influence the virulence phenotype of this strain.

The attenuation of virulence observed in NCTC10247 could not be explained solely by the loss of genes. Only two variable gene clusters were absent from NCTC10247 (groups B and X), but these two groups also were absent from other virulent strains, suggesting that the attenuation may be due to the loss of a single or a few genes rather than a whole cluster. However, pairwise comparisons of each of the six other strains compared with NCTC10247 showed it has lost very few genes compared with any of the strains, and in fact, NCTC10229 had no unique genes relative to NCTC10247. These results were surprising because the other avirulent strains appeared to have lost their virulence through gene loss while cultured in the laboratory. Thus, the mechanism of attenuation is not clear from genomic data and could be due to differential transcriptional control or other reasons such as the disequilibrium of the replichores as discussed above.

The Bm Core Genome Is Under Stronger Purifying Selection Than the Variable Genome To evaluate the evolutionary forces that affect the variable regions in Bm genomes, we constructed detailed alignments and calculated the evolutionary rates for Bm orthologous gene pairs. Significant differences between rates of synonymous (dS) and nonsynonymous (dN) substitutions in the variable and core coding regions of Bm genomes were detected. Both dN and dS values were on average significantly lower for the core gene set: 0.0013 versus 0.0020 for dN ($P = 0.0005$) and 0.0026 versus 0.0033 for dS ($P = 0.0005$). However, the

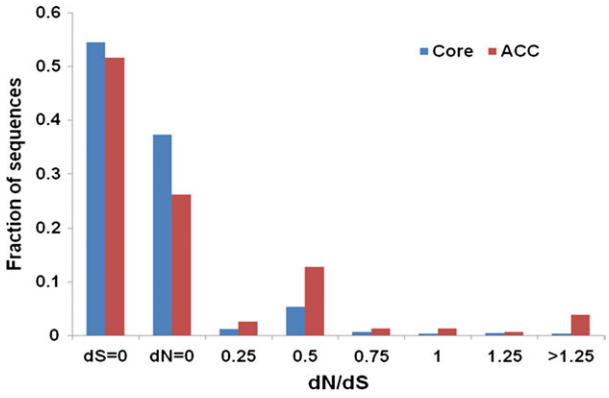

Fig. 6.—Distribution of dN/dS in variable and core genes of Bm genomes aligned with corresponding regions of the reference strain ATCC 23344. dN and dS rates were calculated as described in Materials and Methods. Cumulative data for the seven Bm strains is shown.

selection pressure on nonsynonymous sites varied dramatically between core and variable genes, indicating the existence of stronger purifying selection pressure on Bm core genes. The same trend was observed for virulent strains alone. When three avirulent or attenuated strains were excluded from the analysis, average dN values for the core gene sets were significantly lower than for variable genes ($P < 0.03$). Although overall dN/dS ratios were significantly different between variable and core genes ($P < 0.001$), a large fraction of completely conserved genes (with dS and/or dN equal to 0) was found in both groups (fig. 6) but was lower for variable genes ($P < 0.005$). This trend of higher conservation in the core genes was observed for all individual analyzed strains as well (data not shown), indicating stronger purifying selection on these genes. The observed stronger purifying selection on the core genes is consistent with the hypothesis that the variable genes experience reduced selective pressure within the mammalian host.

Phylogenetic Analysis of Bm An initial phylogenetic analysis comparing the Bm and Bp reference strains relative to nine other *Burkholderia* spp. and to *P. aeruginosa* illustrated the close identity of Bm and Bp (supplementary fig. 2, Supplementary Material online). The two species clustered with the avirulent *B. thailandensis* and were distinct from the other *Burkholderia* spp. as reported previously (Lin et al. 2008). We performed phylogenetic analysis of the Bm species, first using a single nucleotide polymorphism (SNP)-based approach and then by indel analysis. Phylogenetic reconstructions using 515 SNPs as characters indicated that Bm is a monophyletic group and highly consistent with a strictly clonal pattern of evolution (supplementary fig. 3*A*, Supplementary Material online). There were 253 SNPs unique to individual strains and the remaining 262 SNPs defined a highly robust tree with 34 homoplastic SNPs (all nodes had 100% bootstrap support) and a consistency index of 0.84. The root of the tree was determined by polarizing the SNP character states as ancestral or derived by comparison the Bp strain K96243.

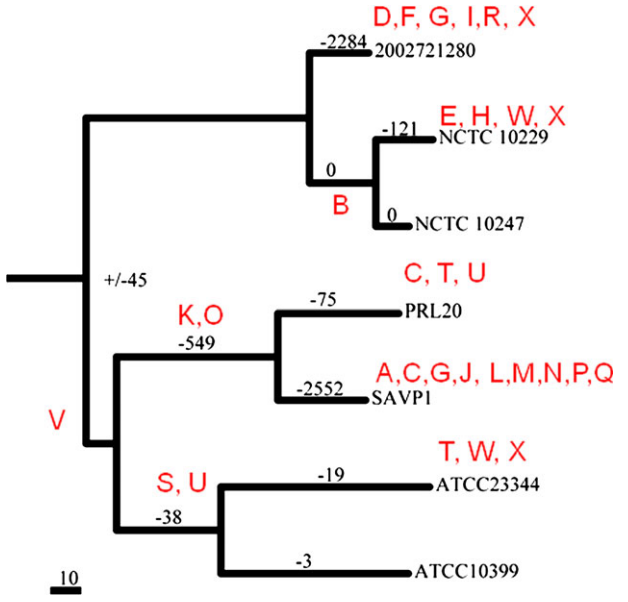

Fig. 7.—Evolutionary tree of Bm showing the number of genes deleted and the evolutionary point of change. In total, 5,686 gene changes can be mapped onto this tree in a manner that assumes only single evolutionary deletion events. Conversely, 997 gene changes require 2 or 3 independent deletions of the same gene. Because we did not compare these genes with Bp, we do not know the ancestral state for 45 of these genes. These 45 genes could be additions or deletions with equal parsimony with mutations occurring along the basal branches of this tree. Letters in red represent the variable regions lost in each branch.

In contrast, indel phylogeny based upon whole-gene differences resulted in a poorly resolved topology and a lower consistency index (0.62; supplementary fig. 3*B*, Supplementary Material online). We found 6,683 genes differing among the seven strains, which was astounding for a recently emerged pathogen. In this analysis, three pairs of highly similar strains clustered together and their association was consistent with the SNP-based tree. The deeper topology, however, was not consistent between the phylogenies. The indel-analysis tree had a four-node polytomy, illustrating the lack of topological resolution.

Different rates of character evolution were clearly seen when gene indels were placed on the SNP-based phylogeny (fig. 7). Some branches had a very large number of gene indels (e.g., 2,284 and 2,552) relative to other branches (0, 3, 45, etc.) of comparable SNP length. Of the 5,683 gene indels analyzed, 997 require two or more "map locations" on the SNP-based tree (data not shown). Superimposing the variable gene cluster data from table 3 revealed that those indels belonged to clusters that had been differentially lost in different strains (fig. 7). The results from the phylogenetic

trees support the hypothesis that Bm evolved from a single Bp ancestor whose genome has been continually rearranged, accompanied by the loss of clusters of genes from different strains in a process of convergent evolution.

Discussion

MLST analysis provided data supporting the evolution of Bm from a single strain of Bp (Godoy et al. 2003). The results presented here from comparative genomic analysis between Bm strains and relative to Bp provide further evidence that Bm arose as a founder population from a single Bp strain, most likely after colonization of an equine-like ancestral host. The evolution of Bm from a Bp ancestor was a result of IS-mediated gene loss and genomic recombination that resulted after genes that provided adaptability to variable environments were no longer under selection in a host. These extraneous genes provided expansion targets for the resident IS element population. Homologous recombination then ensued across IS elements, leading to beneficial genomic losses. Genome evolution continues in Bm, leading to strains that are fitter under different pressures. Our results support the notion that virulence is multifactorial because no gene losses were common among the avirulent and attenuated strains. In addition, the results agree with the hypothesis of genome reduction and erosion as an adaptation to intracellular lifestyle (Ochman and Davalos 2006; Casadevall 2008).

IS element-mediated gene loss in Bm was random and continues to be a major evolutionary mechanism for this species; however, only viable strains can be isolated from an animal host. Random gene loss is evidenced by the unsystematic distribution of variable gene clusters across Bm strains (table 3), and the independent loss of variable clusters in different branches of the phylogenetic tree (fig. 7). In previous laboratory studies, IS407A-mediated gene loss and recombination were observed frequently in vitro (DeShazer et al. 2001; Nierman et al. 2004), and in some cases resulted in lower fitness in an animal host as in SAVP1 (Schutzer et al. 2008) and 200272128. Genomic inversions and rearrangement were a natural outcome of IS expansion with no explicit benefit to Bm, but in some cases, such as NCTC10247, potentially detrimental to the fitness of the organism. This phenomenon of excess IS and other repetitive elements in Bm which mediate recombination and hence rearrangements has been observed in closely related species of other genera, for example, *Bordetella* (Parkhill et al. 2003), *Shigella* (Yang et al. 2005), *Yersenia* (Gu et al. 2007), *Orientia* (Nakayama et al. 2008), and *Clostridium* (Myers et al. 2006), to name a few.

Reconstruction of the ancestral Bp isolate is impractical. First, essentially all the genes in the pan-genome of Bm have already been elucidated (fig. 2), meaning that the closest common ancestor to all Bm strains is most similar to either NCTC10247 or NCTC10399 which harbor the greatest number of variable gene clusters and including those clusters that were lost from each strain (B and X or S, U, and V, respectively). Second, because all the Bp GI have been lost in Bm, it is impossible to infer which, if any, of these GI were present in the ancestor. Sim et al. (2008) found a large number of Bp isolates that have lost all but two of the GIs. Therefore, it is possible that the ancestral Bp strain looked very similar to one of these GI-deficient Bp isolates. Interestingly, those Bp strains were more commonly associated with environmental isolation, rather than human or animal hosts (Sim et al. 2008). Our results from Bm are in better agreement with the findings that there was little correlation between GI content and disease symptoms in melioidosis patients (Tumapa et al. 2008), as all GIs were lost in assuming an obligate mammalian parasite lifestyle. Last, each of the Bm chromosomes has undergone such dramatic rearrangements (fig. 4) that make it almost impossible to discover the ancestral organization of the genome. Although it is possible to conduct a simple concatenation of synteny blocks on known Bp genomes, it is likely that the ancestral Bp strain itself was also rearranged in the process of losing the GIs.

It is noteworthy that the massive intrachromosomal shuffling of gene clusters has occurred with an almost complete absence of interchromosomal recombination. There were no observed interchromosomal exchanges among the any of Bm strains. However, the Bp ancestral strain underwent an interchromosomal exchange that encompassed one of the *rrn* and an anthranilate-resistance operon (fig. 5) in chromosome I. This cluster is located in chromosome II and is flanked by IS407A elements in Bm but not in the Bp genomes. Thus, it is difficult to conclude whether the exchange was induced by IS407A elements that had incorporated into chromosome I of the ancestor or whether the rearrangement sites were hot spots for IS407A insertion after exchange into chromosome II. Interchromosomal *rrn* exchange was observed in *Bartonella* spp. and *Brucella suis* biovar 3 (Jumas-Bilak et al. 1998; Alsmark et al. 2004). However, in both of these genera, the *rrn* exchange occurred from the smaller to the larger replicon, ultimately leading to a reduction in the chromosome number. In Bm, the opposite occurred perhaps as a mechanism to maintain some of the essential genes in the smaller replicon.

Analysis of the shotgun assemblies of other Bp strains revealed that Bp1655 and Bp406e have also undergone dramatic changes in their *rrn* operon content and organization (data not shown). In contrast to Bm that has lost the 3′ end of *rrn* loci, each of these Bp strains has lost the 5′ region of at least one *rrn* locus through recombination across the 23S CDS. These recombinations have resulted in two out of only three major chromosomal rearrangements observed in Bp (Nandi T, et al., in preparation.). Combined, these results suggest that the 23S locus of both Bp and Bm are hot spots

for chromosomal rearrangement and provide further evidence that Bm is the evolutionary product of a single Bp ancestor.

Even though the Bp genome is larger overall, Bm has a larger variable (or accessory) genome. Some explanations are possible for this observation: 1) Bm is an active intermediary step between a "generalist" and an obligate pathogen (Ochman and Davalos 2006; Casadevall 2008). In this case, each Bm strain still carries many genes that granted its Bp ancestor its generalist status. Given enough time, most of these genes would be eroded, resulting in a much smaller genome. 2) The inclusion of reduced virulent strains resulted in an artificially large variable genome because these strains lost IS-defined regions essential for in vivo survival (Schutzer et al. 2008). Certainly, Bm isolated from animals do not resemble SAVP1 or 2002721820 but continuing evolution outside of a host must also account for variability within a species. Compared with Bp, the Bm variable genome functions are not very diverse. Functional role category analysis (table 2) allows us to speculate that Bp has unlimited access to variable genes via lateral transfer and through other means not available to Bm, including phage (Ronning CM, Nierman WC, Ulrich RL, DeShazer D, in preparation). These data show that Bm has entered a population bottleneck and that the small effective population size has further contributed to the homogeneity and reduced genome size of Bm. The resulting Bm population has been at a competitive disadvantage outside of the mammalian host and thus is never isolated from the natural environment.

In summary, our results provide very strong evidence that Bm evolved from a single Bp ancestor through genetic loss and genome rearrangements mediated across IS elements. Bm strains continue to evolve in vivo and in vitro and is another snapshot in our growing understanding of genomic erosion in the path toward adaptation to intracellular lifestyles observed in so many other bacterial pathogens. Further studies into the specific traits lost in avirulent Bm strains, and the potential role of large-scale rearrangements in the reduction of virulence need to be pursued in order to achieve a full understanding of the pathogenicity of Bm.

Funding

This project has been funded with federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services under contract number N01-AI-30071.

Supplementary Material

Supplementary tables 1–3 and supplementary figures 1–3 are available at *Genome Biology and Evolution* online (http://www.oxfordjournals.org/our_journals/gbe/).

Literature Cited

Abbott JC, Aanensen DM, Rutherford K, Butcher S, Spratt BG. 2005. WebACT: an online companion for the Artemis Comparison Tool. Bioinformatics. 21:3665–3666.

Alsmark CM, et al. 2004. The louse-borne human pathogen Bartonella quintana is a genomic derivative of the zoonotic agent Bartonella henselae. Proc Natl Acad Sci U S A. 101:9716–9721.

Atkins T, et al. 2002. A mutant of Burkholderia pseudomallei, auxotrophic in the branched chain amino acid biosynthetic pathway, is attenuated and protective in a murine model of melioidosis. Infect Immun. 70:5290–5294.

Carver T, Thomson N, Bleasby A, Berriman M, Parkhill J. 2009. DNAPlotter: circular and linear interactive genome visualization. Bioinformatics. 25:119.

Casadevall A. 2008. Evolution of intracellular pathogens. Annu Rev Microbiol. 62:19–33.

Chantratita N, et al. 2006. Pulsed-field gel electrophoresis as a discriminatory typing technique for the biothreat agent burkholderia mallei. Am J Trop Med Hyg. 74:345–347.

Dance DA. 1991. Melioidosis: the tip of the iceberg? Clin Microbiol Rev. 4:52–60.

Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Res. 27:4636–4641.

Delcher AL, Phillippy A, Carlton J, Salzberg SL. 2002. Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res. 30:2478–2483.

DeShazer D, Waag DM, Fritz DL, Woods DE. 2001. Identification of a Burkholderia mallei polysaccharide gene cluster by subtractive hybridization and demonstration that the encoded capsule is an essential virulence determinant. Microb Pathog. 30:253–269.

Duerkop BA, et al. 2009. Quorum-sensing control of antibiotic synthesis in Burkholderia thailandensis. J Bacteriol. 191:3909–3918.

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 5:113.

Fleischmann RD, et al. 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science. 269:496–512.

Godoy D, et al. 2003. Multilocus sequence typing and evolutionary relationships among the causative agents of melioidosis and glanders, Burkholderia pseudomallei and Burkholderia mallei. J Clin Microbiol. 41:2068–2079.

Gu J, et al. 2007. Genome evolution and functional divergence in Yersinia. J Exp Zoolog B Mol Dev Evol. 308:37–49.

Haas BJ, et al. 2003. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 31:5654–5666.

Holden MT, et al. 2004. Genomic plasticity of the causative agent of melioidosis, Burkholderia pseudomallei. Proc Natl Acad Sci U S A. 101:14240–14245.

Inglis TJJ, et al. 2000. Interaction between Burkholderia pseudomallei and Acanthamoeba species results in coiling phagocytosis, endamebic bacterial survival, and escape. Infect Immun. 68:1681–1686.

Jumas-Bilak E, Michaux-Charachon S, Bourg G, O'Callaghan D, Ramuz M. 1998. Differences in chromosome number and genome rearrangements in the genus Brucella. Mol Microbiol. 27:99–106.

Kondrashov AS, Shabalina SA. 2002. Classification of common conserved sequences in mammalian intergenic regions. Hum Mol Genet. 11:669–674.

Kothapalli S, et al. 2005. Diversity of genome structure in Salmonella enterica serovar Typhi populations. J Bacteriol. 187:2638–2650.

Kraft C, et al. 2006. Genomic changes during chronic Helicobacter pylori infection. J Bacteriol. 188:249–254.
Lesterlin C, Pages C, Dubarry N, Dasgupta S, Cornet F. 2008. Asymmetry of chromosome Replichores renders the DNA translocase activity of FtsK essential for cell division and cell shape maintenance in *Escherichia coli*. PLoS Genet. 4:e1000288.
Lin CH, Bourque G, Tan P. 2008. A comparative synteny map of Burkholderia species links large-scale genome rearrangements to fine-scale nucleotide variation in prokaryotes. Mol Biol Evol. 25:549–558.
Liu SL, Sanderson KE. 1998. Homologous recombination between rrn operons rearranges the chromosome in host-specialized species of Salmonella. FEMS Microbiol Lett. 164:275–281.
Maiden MCJ, et al. 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U S A. 31:3140–3145.
Moore RA, DeShazer D, Reckseidler S, Weissman A, Woods DE. 1999. Efflux-mediated aminoglycoside and macrolide resistance in Burkholderia pseudomallei. Antimicrob Agents Chemother. 43:465–470.
Myers EW, et al. 2000. A whole-genome assembly of Drosophila. Science. 287:2196–2204.
Myers GS, et al. 2006. Skewed genomic variability in strains of the toxigenic bacterial pathogen, Clostridium perfringens. Genome Res. 16:1031–1040.
Nakayama K, et al. 2008. The whole-genome sequencing of the obligate intracellular bacterium Orientia tsutsugamushi revealed massive gene amplification during reductive genome evolution. DNA Res. 15:185–199.
Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 3:418–426.
Nierman WC, et al. 2004. Structural flexibility in the Burkholderia mallei genome. Proc Natl Acad Sci U S A. 101:14246–14251.
Ochman H, Davalos LM. 2006. The nature and dynamics of bacterial genomes. Science. 311:1730–1733.
Ogurtsov AY, Roytberg MA, Shabalina SA, Kondrashov AS. 2002. OWEN: aligning long collinear regions of genomes. Bioinformatics. 18:1703–1704.
Oliver A, Cantón R, Campo P, Baquero F, Blázquez J. 2000. High frequency of hypermutable Pseudomonas aeruginosa in cystic fibrosis lung infection. Science. 288:1251–1253.
Parkhill J, et al. 2003. Comparative analysis of the genome sequences of Bordetella pertussis, Bordetella parapertussis and Bordetella bronchiseptica. Nat Genet. 35:32–40.
Putman M, van Veen HW, Konings WN. 2000. Molecular properties of bacterial multidrug transporters. Microbiol Mol Biol Rev. 64:672.
Salzberg SL, Delcher AL, Kasif S, White O. 1998. Microbial gene identification using interpolated Markov models. Nucleic Acids Res. 26:544–548.
Sanford JP. 1995. Pseudomonas species (including melioidosis and glanders). In: Mandell GL, Douglas RG Jr., Bennet JE, editors. Principles and practice of infectious diseases. New York: Churchill Livingstone. pp. 1692–1696.
Schmid MB, Roth JR. 1987. Gene location affects expression level in Salmonella typhimurium. J Bacteriol. 169:2872–2875.
Schutzer SE, et al. 2008. Characterization of clinically-attenuated Burkholderia mallei by whole genome sequencing: candidate strain for exclusion from Select Agent lists. PLoS One. 3:e2058.
Shu S, et al. 2000. I-CeuI fragment analysis of the Shigella species: evidence for large-scale chromosome rearrangement in S. dysenteriae and S. flexneri. FEMS Microbiol Lett. 182:93–98.
Sim SH, et al. 2008. The core and accessory genomes of Burkholderia pseudomallei: implications for human melioidosis. PLoS Pathog. 4:e1000178.
Tettelin H, et al. 2005. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proc Natl Acad Sci U S A. 102:13950–13955.
Thibault FM, Hernandez E, Vidal DR, Girardet M, Cavallo JD. 2004. Antibiotic susceptibility of 65 isolates of Burkholderia pseudomallei and Burkholderia mallei to 35 antimicrobial agents. J Antimicrob Chemother. 54:1134–1138.
Toft C, Fares MA. 2008. The evolution of the flagellar assembly pathway in endosymbiotic bacterial genomes. Mol Biol Evol. 25:2069–2076.
Trunck LA, et al. 2009. Molecular basis of rare aminoglycoside susceptibility and pathogenesis of *Burkholderia pseudomallei* clinical isolates from Thailand. PLoS Negl Trop Dis. 3:e519.
Tuanyok A, et al. 2008. Genomic islands from five strains of Burkholderia pseudomallei. BMC Genomics. 9:566.
Tumapa S, et al. 2008. Burkholderia pseudomallei genome plasticity associated with genomic island variation. BMC Genomics. 9:190.
Ulrich RL, DeShazer D. 2004. Type III secretion: a virulence factor delivery system essential for the pathogenicity of Burkholderia mallei. Infect Immun. 72:1150–1154.
Yang F, et al. 2005. Genome dynamics and diversity of Shigella species, the etiologic agents of bacillary dysentery. Nucleic Acids Res. 33:6445–6458.
Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 13:555–556.
Zhou CE, et al. 2007. MvirDB—a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. Nucleic Acids Res. 35:D391–D394.